REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED *(From - To)*
19-11-2012	Final	Sept. 2010 - Mar. 2012

4. TITLE AND SUBTITLE
Reasoning, Learning, And Classifying With Uncertain Causal Models

5a. CONTRACT NUMBER
FA9550-10-1-0466

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)
Rehder, Bob

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Office of Sponsored Programs, New York University
665 Broadway, Suite 801
New York, NY 10012-2331

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
Air Force Office of Scientific Research
875 N. Randolph Street
Arlington, VA 22203
Dr. Myung/RSL

10. SPONSOR/MONITOR'S ACRONYM(S)
AFOSR

11. SPONSOR/MONITOR'S REPORT NUMBER(S)
AFRL-OSR-VA-TR-2012-1242

12. DISTRIBUTION/AVAILABILITY STATEMENT
DISTRIBUTION A: APPROVED FOR PUBLIC RELEASE

13. SUPPLEMENTARY NOTES

14. ABSTRACT
The last 20 years has seen a growing interest in the role of causal knowledge in numerous areas of cognition. One aspect of human causal reasoning studied is how people reason causally under uncertainty, that is, when beliefs are held with less than complete confidence. We address levels of uncertainty are used to resolve how inconsistencies among beliefs are resolved. Another how individual reason with “conjunctive causes”, that is, when a cause only operates when it is enables by other conditions (a spark only yields fire when there is also fuel and oxygen). We propose extensions to causal graphical models (hereafter, CGMs) to represent uncertain causal beliefs and conjunctive causes. Two experiments found that our model of belief integration predicted the qualitative pattern of adults’ causal inferences under uncertainty. It also provided a moderately successful quantitative account of the findings. Another two found that our model predicted the pattern of adults’ inferences with conjunctive causes, with an important exception. A new version of our model extended to account for the exception provided a quite good quantitative account of these experiments.

15. SUBJECT TERMS
Causal reasoning, causal learning, belief integration, belief revision, uncertainty, interactive causes, conjunctive causes, independent causes

16. SECURITY CLASSIFICATION OF: a. REPORT	b. ABSTRACT	c. THIS PAGE	17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
					19b. TELEPHONE NUMBER *(Include area code)*

Reset

INSTRUCTIONS FOR COMPLETING SF 298

1. REPORT DATE. Full publication date, including day, month, if available. Must cite at least the year and be Year 2000 compliant, e.g. 30-06-1998; xx-06-1998; xx-xx-1998.

2. REPORT TYPE. State the type of report, such as final, technical, interim, memorandum, master's thesis, progress, quarterly, research, special, group study, etc.

3. DATES COVERED. Indicate the time during which the work was performed and the report was written, e.g., Jun 1997 - Jun 1998; 1-10 Jun 1996; May - Nov 1998; Nov 1998.

4. TITLE. Enter title and subtitle with volume number and part number, if applicable. On classified documents, enter the title classification in parentheses.

5a. CONTRACT NUMBER. Enter all contract numbers as they appear in the report, e.g. F33615-86-C-5169.

5b. GRANT NUMBER. Enter all grant numbers as they appear in the report, e.g. AFOSR-82-1234.

5c. PROGRAM ELEMENT NUMBER. Enter all program element numbers as they appear in the report, e.g. 61101A.

5d. PROJECT NUMBER. Enter all project numbers as they appear in the report, e.g. 1F665702D1257; ILIR.

5e. TASK NUMBER. Enter all task numbers as they appear in the report, e.g. 05; RF0330201; T4112.

5f. WORK UNIT NUMBER. Enter all work unit numbers as they appear in the report, e.g. 001; AFAPL30480105.

6. AUTHOR(S). Enter name(s) of person(s) responsible for writing the report, performing the research, or credited with the content of the report. The form of entry is the last name, first name, middle initial, and additional qualifiers separated by commas, e.g. Smith, Richard, J, Jr.

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES). Self-explanatory.

8. PERFORMING ORGANIZATION REPORT NUMBER. Enter all unique alphanumeric report numbers assigned by the performing organization, e.g. BRL-1234; AFWL-TR-85-4017-Vol-21-PT-2.

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES). Enter the name and address of the organization(s) financially responsible for and monitoring the work.

10. SPONSOR/MONITOR'S ACRONYM(S). Enter, if available, e.g. BRL, ARDEC, NADC.

11. SPONSOR/MONITOR'S REPORT NUMBER(S). Enter report number as assigned by the sponsoring/ monitoring agency, if available, e.g. BRL-TR-829; -215.

12. DISTRIBUTION/AVAILABILITY STATEMENT. Use agency-mandated availability statements to indicate the public availability or distribution limitations of the report. If additional limitations/ restrictions or special markings are indicated, follow agency authorization procedures, e.g. RD/FRD, PROPIN, ITAR, etc. Include copyright information.

13. SUPPLEMENTARY NOTES. Enter information not included elsewhere such as: prepared in cooperation with; translation of; report supersedes; old edition number, etc.

14. ABSTRACT. A brief (approximately 200 words) factual summary of the most significant information.

15. SUBJECT TERMS. Key words or phrases identifying major concepts in the report.

16. SECURITY CLASSIFICATION. Enter security classification in accordance with security classification regulations, e.g. U, C, S, etc. If this form contains classified information, stamp classification level on the top and bottom of this page.

17. LIMITATION OF ABSTRACT. This block must be completed to assign a distribution limitation to the abstract. Enter UU (Unclassified Unlimited) or SAR (Same as Report). An entry in this block is necessary if the abstract is to be limited.

Reasoning, Learning, and Classifying with Uncertain Causal Models

Grant Award No. FA9550-10-1-0466
Jay Myung, PhD

Bob Rehder
New York University
November 19, 2012

0. Preamble and Introduction

The last 20 years has seen a growing interest in the role of causal knowledge in numerous areas of cognition. Many studies have investigated how causal relations are learned from observed correlations (Cheng, 1997; Gopnik et al., 2004; Griffiths & Tenenbaum, 2005; 2009; Lu et al., 2008; Sobel et al., 2004; Waldmann et al., 1995). Others have tested their impact on various forms of reasoning, including inference (Kemp & Tenenbaum, 2009; Kemp et al., 2012; Oppenheimer, 2004; Rehder, 2006; 2009; Rehder & Burnett, 2005), interventions (Sloman & Lagnado, 2005; Waldmann & Hagmayer, 2005), decision making (Hagmayer & Sloman, 2009), analogy (Holyoak et al., 2010; Lee & Holyoak), and classification (Rehder & Hastie, 2001; Rehder 2003a; b; Rehder & Kim, 2006; 2009; 2010).

The original goal of this proposal (intended to cover 3 years of research but funded for 18 months) was to study three aspects of human causal reasoning. The first is how people reason causally under uncertainty, that is, when beliefs are held with less than complete confidence. We address a particular application in which a representation of uncertainty resolves how inconsistencies among beliefs are resolved. The second aspect is how individual reason with "conjunctive causes", that is, when a cause only operates when it is accompanied by one or more other causes (a spark only yields fire when there is also fuel and oxygen). The third project concerns how people reason with inhibitory causes, that is, factors that disable or deactivate a causal mechanism. Research on the first two of these topics was conducted and is near completion, as now described.

1. Scientific Objectives of Research

The aim of this research was to advance both empirical knowledge about and theoretical accounts of human causal reasoning. One topic concerned how people reason causally under uncertainty. We take it as a given that knowledge comes to us in many forms. Much of what we know about the world comes from what we are told by others, but these sources can vary greatly in how informed (and trustworthy) they are and the effectiveness with which they communicate their message. Avoiding second-hand sources by observing things for oneself is important but has its own drawbacks: observations are always finite in number, are often incomplete (not all variables are measured), and are susceptible to measurement error (due, e.g., to failures of perception and memory). In many scientific domains, direct observation is impossible for all but highly trained experts. Finally, people may have default expectations (i.e., "priors') regarding, e.g., the strength and number of causal relationships. These facts mean that people have multiple sources of knowledge that vary in their reliability, format, and completeness. In this light, it is inevitable that inconsistencies among those sources will arise such in that they cannot be reconciled into a single coherent theory of the domain. Yet, people must draw inferences nevertheless. How do they do so?

The second topic concerns how people reason in light of the fact that causes usually do not operate in a vacuum but rather interact with other factors to produce their effects. E.g., the conjunction of two or more variables is often necessary for an outcome to occur. A spark may only produce fire if there is fuel to ignite, a virus may only cause disease if one's immune system is suppressed, the motive to commit murder may result in death only if the means to carry out the crime are available. Sometimes, conjunctive causes take the form of *enablers*. E.g., the presence of oxygen enables fire given spark and fuel. Although some studies have investigated the *learning* of interactive causes (e.g., Novick & Cheng, 2004), in this research we examine their role in *reasoning*.

These projects each include both a theoretical and empirical component. Theoretically, we operate within a probabilistic framework that has become popular for modelling learning and reasoning with causal knowledge, namely, *Bayesian networks* or *causal graphical models* (hereafter, *CGMs*). We propose extensions to represent uncertain causal beliefs and conjunctive causes. From these extensions we derive how causal inferences should be made, predictions that are tested by assessing how people reason under a variety of experimental conditions. We also assess our model's account by quantitative fitting our models to peoples' inferences.

2. Technical Approach

Causal Reasoning Under Uncertainty: A Theoretical Model of Belief Integration

As argued, realistic causal reasoning involves inferences with individual beliefs that might be inconsistent with one another. On one hand, there is evidence from social psychology showing people are often untroubled by contradictory beliefs. Nevertheless, inconsistent beliefs might all contribute to an *inference*. We propose that reasoning can be modeled by a two-step process in which beliefs are first integrated into a consistent "causal model" that is then used to compute the inference. Modeling the process of integration in turn requires specifying the *confidence* with which beliefs are held and then how they are made consistent.

Representing causal uncertainty. We assume that uncertainties are represented in the form of probability density functions on each component of the causal model. We consider two fundamental types of knowledge: the probability that variables take on different values and the strength (or "power") of causal relationships that relate them. For simplicity, we only consider binary variables and generative causal links that operate independently. Let $\mathbf{V}$ represent the set of variables in the domain. For each $v \in \mathbf{V}$, assume that v occurs (is "present" rather than "absent") with probability π^v and that π^v is a beta distribution that assigns a subjective degree of belief to every value in the range [0–1]. Next, let $\mathbf{L}$ be the set of explicit causal links in a model. For each $l \in \mathbf{L}$, l's causal power (the probability that it operates when the cause is present) is characterized by a beta distribution π^l. Finally, let $\mathbf{E}$ be the subset of variables in $\mathbf{V}$ that are effects. Each $e \in \mathbf{E}$ is assumed to potentially have causes that are not explicit in the model; the influence of these alterative or "background" causes are aggregated into a single causal link. The power of this background cause is a beta distribution π^e. Each beta distribution π

has an expected value of $E[\pi] = \alpha/[\alpha + \beta]$. The sum of α and β, referred to here as $f(\pi)$, can be interpreted as the confidence with which the reasoner believes in $E[\pi]$.

Integrating causal beliefs. Let ***r***, ***m***, and ***b*** be vectors specifying the base rate of every variable in **V**, the strength of every link in **L**, and the strength of the background causes of every effect in **E**. A consistent model is one in which effects are neither *under*- nor *over-determined*, i.e., their base rates are exactly explained by their causes. For independent, generative causes, this constraint can be expressed by a "fuzzy or" equation relating the probability of an effect *e* to its parents,

$$r_e = 1 - (1 - b_e)\left[\prod_{l \in \mathbf{L} \& l.e = e} (1 - r_{l.c} m_l)\right] \tag{1}$$

where *l.c* and *l.e* are the cause and effect variables associated with causal link *l*, so that the product ranges over all the causal links between *e* and its parents.

We stipulate that the joint probability distribution on the parameters of consistent models can be formed from the parameters' individual π distributions, with those parameters that imply under- and over-determination of any effect variable assigned a probability of 0,

$$\begin{aligned} p(\boldsymbol{r}, \boldsymbol{m}, \boldsymbol{b}) \quad &= 0 \text{ when } \exists_{e \in \mathbf{E}} r_e \neq 1 - (1 - b_e)\left[\prod_{l \in \mathbf{L} \& l.e = e} (1 - r_{l.c} m_l)\right] \\ &\propto \prod_{v \in \mathbf{V}} \pi^v(r_v) \prod_{l \in \mathbf{L}} \pi^l(m_l) \prod_{e \in \mathbf{E}} \pi^e(b_e) \text{ otherwise.} \end{aligned} \tag{2}$$

The most likely causal model parameters that resolve potential inconsistencies among causal beliefs are the ***r***, ***m***, and ***b*** that maximize Eq. 2.

Testing the model. We created experimental analogs of situations we believe are common: one has some theoretical beliefs (i.e., causal laws) and some statistical knowledge (e.g., facts about the base rates of events) and these beliefs are mutually inconsistent. Eq. 2 works to ensure that reasoners' inferences reflect consistent beliefs. E.g., the most likely base rate for variable *v* will reflect not only what one has knows about *v*, but also the strengths of the causal relationships in which *v* is involved (and the base rates of other variables to which *v* is causally related). Conversely, the most likely strength for causal link *l* will reflect not only what we know about *l* but also what is known about the base rates of its cause and effect (and the strengths of causal links in which they are involved). These claims are tested in two experiments assessing adults' causal inferences.

Reasoning With Conjunctive Causes

As argued, causes usually only produce their effects when conjoined with other factors, such as enabling conditions. Fig. 1A presents a CGM in which variables C_1 and C_2 are causes of variable E. By itself, however, this model says nothing about the functional relationship between E and its causes. Fig. 1B represents the fact that C_1 and C_2 are independent causes of E—that is, that E might be caused by C_1 *or* C_2. Fig. 1C represents that C_1 and C_2 are conjunctive causes of E—E is brought about only when C_1 *and* C_2 are both present.

We tested how adults reason with independent vs. conjunctive causes. To assess our model we adopted the novel methodology of asking people to make three distinct types of inferences, namely, judgments of *joint*, *conditional*, and *marginal* probability. We specify the joint probability, from which the other types of judgments can be derived. In general, we have

$$p(C_1,C_2,E) = p(E \mid C_1,C_2)\,p(C_1,C_2) \qquad (3)$$

Assuming that C_1 and C_2 have no hidden common causes, Eq. 3 becomes

$$p(C_1,C_2,E) = p(E \mid C_1,C_2)\,p(C_1)\,p(C_2) \qquad (4)$$

$p(E|C_1,C_2)$ can be written as a function of parameters that characterize the generative causal mechanisms that relate E to its causes. For independent causes,

$$p(E \mid C_1,C_2) = 1-(1-b_E)\prod_{i=1,2}\left(1-m_{C_i,E}\right)^{ind(C_i)} \qquad (5)$$

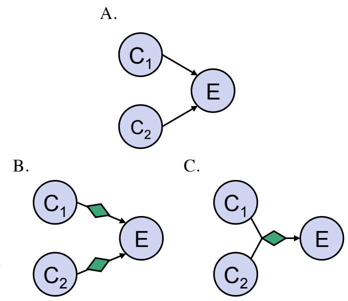

Figure 1.

where $m_{C_1,E}$ and $m_{C_2,E}$ represent the probabilities that those mechanisms will produce E when C_1 and C_2 are present, respectively, b_E represent the probability that E will be brought about by one or more additional causes not shown in Fig. 1, and $ind(C_i)$ returns 1 when C_i is present and 0 otherwise. For conjunctive causes,

$$p(E \mid C_1,C_2) = 1-(1-b_E)\left(1-m_{C_1,C_2,E}\right)^{ind(C_1,C_2)} \qquad (6)$$

where $m_{C_1,C_2,E}$ is the probability that C_1 and C_2 will bring about E when both are present and $ind(C_1, C_2)$ returns 1 when C_1 and C_2 are both present and 0 otherwise.

Testing the model. We instructed subjects on scenarios involving both independent and conjunctive causes and asked them to make judgments of joint, conditional, and marginal probability. Expt. 1 directly compared subjects' inferences with independent and conjunctive causes. Expt. 2 tested how the pattern of inferences changes as a function of the strength of the causal relations.

Another goal was to assess proposals regarding how people often augment causal representations with additional knowledge. Rehder & Burnett (2005) found that people's causal inferences about category features exhibited a systematic deviation from the predictions of CGMs in which a feature was rated as more likely when other features were present, even those features were (according to the Markov condition) conditionally independent. Rehder & Burnett suggested that this *typicality effect* arises because people assume that category features are related via hidden causal mechanisms. On this account, we should also see similar effect in the present experiments, which also tested category features.

3. Progress Made & Results Obtained

Causal Reasoning Under Uncertainty

The proposed model of causal uncertainty and belief integration was tested in two experiments by providing subjects with multiple sources of both empirical and theoretical information. Each experiment used a 2 x 2 design in which one factor varied the base rates of the causes and the other the strength (Expt. 1) or number (Expt. 2) of causal links. Manipulations of these two types of causal model parameters should result in changes to not only the parameters themselves but also to the other parameters in the model in order to achieve model consistency.

Due to space limitation, only Expt. 2. is presented here (see McDonnell et al., in preparation, for a complete report of both experiments). Subjects learned about 4 binary variables (referred to here as C_1, C_2, C_3, and E) in the domains of economics, meteorology, or sociology. E was described as having one cause (C_1) or three (C_1, C_2, and C_3). The cause variables were described as being either rare (appearing in 25% of instances) or common (75%). All subjects were told that the effect E was "somewhat common," occurring with probability 44%, and that it had no other causes besides those on which they were instructed. These knowledge sources were described as reflecting the "beliefs," "estimates," and "tentative theories" of domain experts rather than established facts. Subjects were then asked to make a number of judgments of both *conditional* (predict one variable from others) and *joint* (the likelihood of a particular set of variables) probability.

In some cells of this experiment the supplied information was inconsistent. E.g., the rare/1-link condition is inconsistent because, according to Eq. 1, rare causes and one causal link (and the absence of any other causes of E) imply a base rate for E of only .125, as compared to the instructed base rate of .44. To compensate, subjects' inferences should thus reflect an updated model with causes that are more prevalent, stronger causal links, an effect that is less prevalent, or some combination of all of these. Conversely, in the common/3-link condition the effect is over-determined because common causes and strong causal links imply a base rate for E of .76 as compared to .44. To compensate, subjects' inferences should reflect a model with causes that are less prevalent, causal links that are weaker, an effect that is more prevalent, or some combination of all of these.

We also asked whether subjects would update their causal model in light of data they observe. To this end, after making inferences on the basis of the instructed information, subjects observed samples of 32 economies (or weather systems or societies). We refer the reader to McDonnell et al. for these results.

Results. To characterize subjects' causal model, their ratings of both joint and conditional probability in each block were fit to a causal model with one effect and three independent and generative causes. These fits yielded a total of 8 causal model parameters per subject per block: r_{C_1}, r_{C_2}, r_{C_3}, m_1, m_2, m_3, b, and r_E. For notational convenience we refer to r_{C_i} and r_E as c_i and e, respectively.

With subjects' causal models thus characterized, we now turn to the central question of whether those models exhibit the predicted pattern of integration of inconsistent knowledge sources. Fig. 2 presents the effects of our two manipulations on subjects' fitted causal model parameters. (It also includes the fits of our theoretical model, discussed below.) For comparison, we only present those parameters that were involved in a causal link in all conditions (i.e., c_1, m_1, e, and b). First, a main effect of causal bases rates in Fig. 2A confirmed the success of that manipulation. Importantly, this manipulation also increased the prevalence of the effect (parameter e in Fig. 2C) and decreased the strength of the background causes (parameter b in Fig. 2B). That is, to accommodate the greater prevalence of the causes, subjects compensated by increasing the base rate of the effect and lowering the strength of the alternative causes. These changes are examples of the types of responses predicted by our theoretical model of belief integration. There was no effect of the causes' base rates on causal strengths (parameter m_1 in Fig. 2B).

The manipulation of the number of causal links also had two important effects. First, it reduced the base rate of C_1 (parameter c_1 in Fig. 2A). Second, it reduced the strength of the $C_1 \rightarrow E$ causal relationship (parameter m_1 in Fig. 2B). That is, to accommodate two additional causal links, subjects compensated by decreasing the effectiveness of the $C_1 \rightarrow E$ causal link (by reducing both c_1 and m_1). Again, these changes were predicted by our model as being necessary to obtain belief consistency. The number of causal links affected neither the prevalence of the effect nor the strength of alternative causes (parameters *e* and *b*).

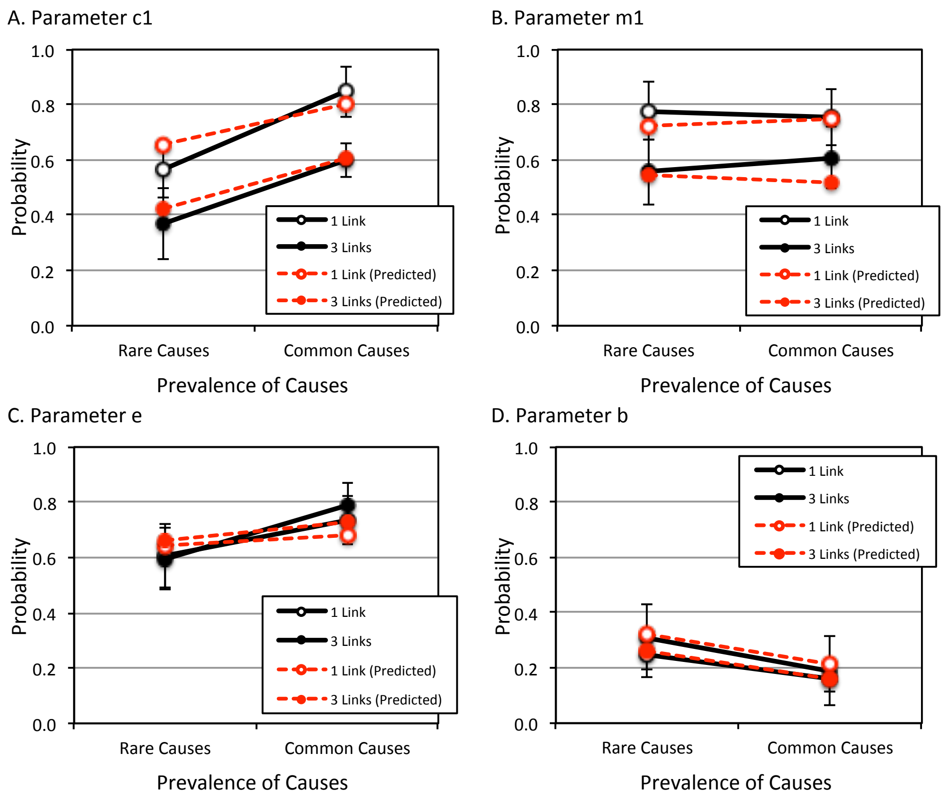

Fig. 2. Causal model parameters derived from the block 0 judgments (i.e., before any samples were observed). (A) The base rates of C_1 (parameter c_1). (B) The strength of the causal links (m_1). (C) The base rate of the effect (e). (D) The strength of alternative causes (b). Error bars are standard errors. The fits of our belief integration model are show in red.

Theoretical modeling. We also assessed our theoretical model by fitting it to the results in Fig. 2. Recall that although it specifies that reasoners will adjust their beliefs so as to reason with a consistent model, it does not specify which beliefs should be adjusted in the absence of any information about the confidence with they are held. Accordingly, the confidence that reasoners have in each instructed model component is represented as four free parameters: f^c (the base rates of the cause), f^m (the strength of the causal links), f^b (the strength of alternative causes), and f^e (the base rate of the effect).

In addition, because we assume that subjects do not perfectly encode the initial numerical information provided about each model component, those are free parameters as well. Let parameters in^m, in^b, and in^e represent the instructed values of *m* (causal link strengths, described to subjects as .50), *b* (the strength of alternative causes, described as 0), and *e* (the base rate of the effect, described as

.44), respectively. in^m was the initial strength of all three links in the 3-link conditions and $C_1 \rightarrow E$ in the 1-link conditions; otherwise, the initial strength was set to 0. Parameters $in^{c\text{-}rare}$ and $in^{c\text{-}common}$ represent the initial base rates of the causes in the rare and common conditions, respectively (described to subjects as .25 and .75). Subjects' judgments in each block were fit assuming each π distribution was updated to reflect the observed samples.

This model was fit to the group level causal model parameters in Expt. 2, (i.e., the 8 parameters [c_1, c_2, c_3, m_1, m_2, m_3, b, and e] estimated per block per condition). The parameters that minimized squared error were f^c = 140, f^m = 156, f^b = 323, f^e = 845, $in^{c\text{-}rare}$ = .203, $in^{c\text{-}common}$ = .495, in^m = .495, in^b = .118, and in^e = .799. The correlation between the "observed" and predicted values was .961. The causal model parameter values generated by this fit are presented in Fig. 2.

Fig. 2 reveals that the model is able to capture each of the effects shown in Fig. 2. The base rate of C_1 (parameter c_1) and the strength of the $C_1 \rightarrow E$ link (m_1) both decrease as the number of causal relations increase (Figs. 2A and B). And, the base rate of the effect (e) increases and the decrease in alternative causes (b) decreases as the base rate of the causes increases (Figs. 2C and D), although e's increase is clearly smaller in magnitude than that exhibited by subjects. The insensitivity of e to changes in the other model components is a manifestation of the relatively large confidence placed in its initial value (f^e = 845 vs. all other fs < 400).

Discussion. These result provide initial support for our claim that causal inferences are made by first integrating multiple, inconsistent beliefs into a coherent domain theory. Much more needs to be done of course. If our account is correct, then inferences should be affected by manipulations of confidence (low confidence parameters should undergo more changes than high ones) and indeed this is our next experiment. And, although the fitting results shows that our models hold promise as an account for these sort of data, the fact that subjects were relatively insensitive to the data they observed reduces the effective number of data points, raising the specter of overfitting. Again, experiments are planned to increase subjects' sensitivity to data and thus provide a more stringent test of the model.

Reasoning With Conjunctive Causes

The predictions of the joint, conditional, and marginal probability derived from CGMs for independent vs. conjunctive causes were tested in two experiments. As mentioned, Expt. 1 compared subjects' inferences with independent vs. conjunctive causes (and whether those differences corresponded to the predictions presented above) whereas Expt. 2 assessed how they are affected by the strength of the causal relations. Due space limitations only the results of Expt. 1. are presented here (see Rehder, in preparation, for a complete report of Expts. 1 and 2). In Expt. 1, subjects were taught novel categories with 6 features. Within each category, each of the feature triplets was described as forming either an independent (Fig. 1B) or conjunctive (Fig. 1C) network. Subjects were then presented with marginal, joint, and conditional probability judgments. 48 NYU undergraduates served as subjects.

Results. Feature inference (i.e., conditional probability) ratings are presented in Figs. 3A-C. Fig. 3A presents ratings regarding the effect given the presence of 0, 1, or 2 of the causes. Unsurprisingly, subjects judged that the presence of the effect was rated to be very likely (ratings > 90) when both causes were present and very unlikely (< 15) when both were absent for both types of causal networks. But when just one cause was present, subjects were much more likely to predict the effect for the independent (rating of 80) as compared to conjunctive (27) network. This result confirms that subjects are sensitive to the form of the functional relationship that relates effects and their causes.

Fig. 3A also reveals a way that the ratings differed from the predictions. Subjects judged that the conjunctive effect was more probable in the presence of one cause (27) vs. none (11) when in fact those ratings should be equal. This result corresponds to the typicality effect described earlier in which features that should be conditionally independent are dependent instead (Rehder & Burnett, 2005).

Fig. 3B shows ratings when subjects predicted a cause given the presence of the effect, as a function of whether the *other* cause was present. When causes were independent, the cause was rated higher when the other cause was absent (84) vs. present (70), reflecting the well-known *explaining away* phenomenon in which the presence of one cause that accounts for an effect makes other causes less likely. In contrast, this pattern was reversed for conjunctive causes (58 vs. 94). We refer to this phenomenon as *exoneration*. E.g., murder requires not only the motive but also the means, so discovering that a murder suspect didn't possess the means to carry out the crime (e.g., proximity to the victim) decreases his likely guilt.

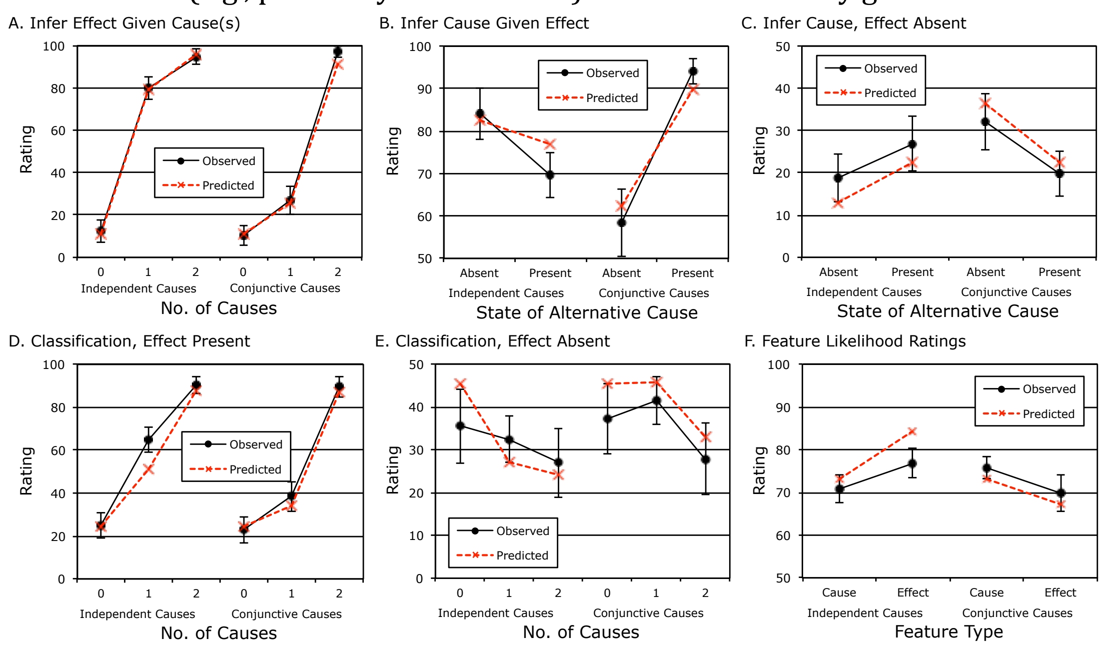

Fig. 3. Empirical results from Expt. 1. Inference (conditional probability) judgments are in panels A-C, classification judgments (joint probability) are in panels D-E, and feature likelihood (marginal probability) judgments are in panel F. Error bars are 95% confidence intervals. Fits of the underlying mechanism model are superimposed in red on the observed data.

Finally, Fig. 3C shows subjects' prediction of a cause given the absence of the effect. A conjunctive cause was rated higher when the other conjunct was absent (32 vs. 20 when present), reflecting another kind of exoneration: Your sister, who failed to show up for Thanksgiving, is exonerated when you learn that her flight (the enabler) was canceled. In contrast, independent causes exhibit a typicality effect in which a cause was rated as more likely when the other cause was present (27 vs. 19) even though those causes were supposed to be conditionally independent.

Classification (i.e., joint probability) ratings are presented in Figs. 3D and 3E. When the effect was present (Fig. 3D) objects received low ratings when both causes were absent and high ones when they were both present, as expected. But ratings were much higher when one cause was present for independent vs. conjunctive causes (63 vs. 39). When the effect was absent (Fig. 3E), ratings *decreased* as the number of independent causes increased. Rehder & Kim (2006; 2010) refer to this phenomenon as a *coherence effect* in which an object is a better category member when its features corroborate the category's causal links (cause and effect features both present or both absent) and a worse one when they violate those links (cause present and effect absent or vice versa). For conjunctive causes, ratings reflected the predicted nonmonotonic change in ratings as the number of causes increased, with the item with one cause receiving the highest ratings.

The feature likelihood (i.e., marginal probability) ratings in Fig. 3F exhibit the expected interaction between network type and feature type: an effect was rated as less probable when it was conjunctive rather than independent (66 vs. 77).

Theoretical modeling. To assess whether the independent and conjunctive CGMs in Fig. 1 also provide a quantitative account for these results, they were simultaneously fit to the inference, classification, and feature likelihood ratings. Because it is known a priori that those models will be unable to account for the typicality effect, we followed Rehder & Burnett (2005) and augmented them with a node representing an underlying mechanism (UM), as shown in Fig. 4. The *c* parameters associated with the four explicit causes in Fig. 2 were assumed to be equal as were all *m* parameters and all *b* parameters. c_{UM} is the probability that UM is present and m_{UM} is the power of the causal links between it and the 6 category features. A γ parameter for each judgment type applied a nonlinear power transformation of the probability derived the CGMs onto subjects' ratings.

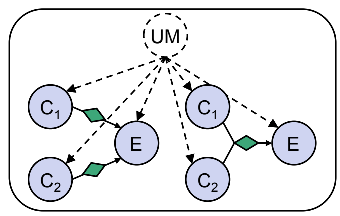

Figure 4.

This model was fit to each subject's 30 ratings with parameters that minimized squared error. The best fitting model parameters averaged over subjects were c = .505, m = .704, b = .148, c_{UM} = .609, and m_{UM} = .763. The predictions are presented in Fig. 3 superimposed on the empirical data. The figure shows that the model was able to account for most of the qualitative effects seen in this experiment, including explaining away and exoneration effects in inference, coherence effects in classification, and the marginal probabilities of the independent and conjunctive effects. Importantly, it also accounted for the typicality effect (e.g., the larger ratings when predicting a conjunctive effect given one vs. zero causes; Fig. 3A).

Of course, Fig. 3 also reveals some mispredictions. E.g., the model predicts a smaller explaining away effect than exhibited by subjects (Fig. 3B) and overpredicts the marginal probability of the independent cause (Fig. 3F). Nevertheless, the model's ability to account for the effects in Fig. 3 resulted in a correlation between the observed and predicted ratings of .984 (.895 averaged over subjects).

4. Significance of Results & Impact on Science

Integrating Causal Beliefs

Most studies of causal reasoning ask subjects to reason in simplified condition involving a domain theory that is simple and free of inconsistencies. While these findings are valuable, realistic causal reasoning usually take place in the context of multiple sources of possibly contradictory information. We created experimental analogs of a situation in which one has some theoretical beliefs (e.g., candidate causal laws) and some elementary statistical knowledge (e.g., facts about the base rates of events) and the later are inconsistent with the former. Our subjects' inferences reflected the sort of changes needed to reason with a coherent causal theory of the domain. Making causes more prevalent resulted in alternative causes becoming weaker and the effect becoming more prevalent. Making causal relations more numerous resulted in the causes becoming rarer and other causal links becoming weaker. We know of *no* other model that is capable of predicting these sorts of effects. It also provided a moderately successful quantitative account of these processes.

Our experiments did not exhaust the types of knowledge that may influence a causal inference of course. We presented statistical information about variables' base rates in both verbal form and samples of observations, but real-world observations are often incomplete in that some variables have missing values. Correlational information might come to us in verbal form rather than via observations (e.g., "mental illness and homelessness tend to go together"). And so on. The challenge faced by theorists then is to specify how these many different kinds of knowledge sources are integrated. We believe that our model represents the start of a solution to this problem.

Finally, our model may have much to say about not just how beliefs are integrated but also how they are *revised*, that is, permanently changed in light of counterevidence. Although social psychology tells us it happens rarely, people do occasionally revise their beliefs, and we suspect that the basic factors embodied in our model are some of the precursors. E.g., beliefs that are highly inconsistent with others and thus need to be greatly modified during the integration process may eligible for more permanent revisions. More extreme revisions might involve not only change to parameters but also structural changes in which causal relations are added or deleted from the model. In this manner, the integration of contradictory beliefs becomes another factor that determines how one chooses between "causal hypotheses" (i.e., alternative causal structures) (Griffiths & Tenenbaum, 2005).

Causal Reasoning Under Uncertainty

This work opens new avenues of research into how people acquire, represent, and use their beliefs about the reliability of acquired knowledge. Although our model of uncertainty was developed to specify how beliefs are integrated, it is clear that virtually all causal inference will be affected by the confidence with which beliefs are held. E.g., all else being equal, one will judge that C_1 is more likely than C_2 to be responsible for E if the $C_1 \rightarrow E$ causal link is held with more confidence than $C_2 \rightarrow E$; for the same reason, if one wants to intervene to bring about E one should manipulate C_1 rather than C_2. One will infer an effect from a cause more certainly when the enabling conditions necessary for the causal mechanism to operate are confidently assumed to be present. And so forth.

Although our representation of uncertainty was sufficient to account for our empirical results, its assumption that the distributions of the individual causal model parameters are independent is certain to be unrealistic in some situations. E.g., Lu et al. (2008) have modeled the traditional causal learning experiment as one in which the prior distribution is a two-dimensional density function on (a) the strength of the to-be-learned causal link (parameter *m* in our terms) and (b) the strength of alternative causes (parameter *b*). When density is massed around either large *m*/small *b* or small *m*/large *b*, this model captures subjects' apparent preferences for a single causal explanation of the effect. (By so doing, it reproduces a number of the effects that Griffiths & Tenenbaum, 2005, accounted for in terms of a hypothesis-testing model; also see Lombrozo, 2007.) Disjunctive hypotheses of this sort might take other forms. E.g., it may be conveyed through explicit instruction ("either mental illness *or* unemployment is the cause of homelessness").

Reasoning with Conjunctive Causes

One question asked in this research is whether causal inferences are sensitive to the functional relationship that can relate an effect to its causes. The answer is that they are: judgments of conditional, joint, and marginal probability all differed depending on whether causes were described as independent or conjunctive. Moreover, those inferences exhibited the patterns predicted by a generative representation of causal knowledge. When causes were independent, subjects exhibited explaining away, an expected result given past demonstrations of that effect in the social and cognitive literatures. But when causes were conjunctive, subjects exhibited what we referred to as exoneration effects. To our knowledge, this is the first demonstration that exoneration effects are entailed by conjunctions causes and that human causal reasoners in fact exhibit that effect.

Other frameworks for representing causal knowledge are unable to readily explain these results. E.g., simple spreading activation networks are clearly unable to account for the present result because such networks are insensitive to both the distinction between independent and conjunctive causes. Proposals that treat such knowledge as a dependency network (Sloman et al., 1998) is sensitive to causal direction (E "depends on" C_1 and C_2) but is also unable to explain effects that depend on the nature of the functional relationship between causes and effects.

Our experiments also replicated previous research showing that people's inferences exhibit a violation of conditional independence known as the typicality effect, and extended this effect to conjunctive causes. Model fitting showed CGMs augmented with an additional "underlying mechanism" were able to almost fully account for the inferences generated for both independent and conjunctive causes.

Finally, we know of no other study that has tested models of causal reasoning using *multiple* types of judgments (in our cause judgments of marginal, conditional, and joint probability). This methodological innovation is important because of well known problems associated with establishing for the psychological reality of mental representations. Going forward, we believe that accounting for these sorts of "converging operations" (i.e., multiple types of judgments) should be the standard against which theories of causal inference should be held.

5. Publications Resulted from Research

Rehder, B. (2011). Reasoning with conjunctive causes. In L. Carlson. C. Hoelscher, & T.F. Shipley (Eds.), *Proceedings of the 33rd Annual Conference of the Cognitive Science Society* (pp. 1406-1411). Austin, TX: Cognitive Science Society.

6. References Cited

Cheng, P. (1997). From covariation to causation: A causal power theory. *Psychological Review, 104*, 367-405.

Gelman, S. A. (2003). The essential child: The origins of essentialism in everyday thought. New York: Oxford University Press.

Gopnik, A., Glymour, C., Sobel, D. M., Schulz, L. E., & Kushnir, T. (2004). A theory of causal learning in children: Causal maps and Bayes nets. *Psychological Review, 111*, 3-23.

Griffiths, T. L., & Tenebaum, J. B. (2005). Structure and strength in causal induction. *Cognitive Psychology, 51*, 334-384.

Griffiths, T. L., & Tenenbaum, J. B. (2009). Theory-based causal induction *Psychological Review, 116*, 56.

Hadjichristidis, C., Sloman, S. A., Stevenson, R., & Over, D. (2004). Feature centrality and property induction. *Cognitive Science, 28*, 45-74.

Hagmayer, Y., & Sloman, S. A. (2009). Decision makers conceive of themselves as interveners. *Journal of Experimental Psychology: General, 128*, 22-38.

Holyoak, K. J., Lee, J. S., & Lu, H. (2010). Analogical and category-based inferences: A theoretical integration with Bayesian causal models. *Journal of Experimental Psychology: General, 139*, 702-727.

Kemp, C., & Tenenbaum, J. B. (2009). Structured statistical models of inductive reasoning. *Psychological Review, 116*, 20-58.

Kemp, C., Shafto, P., & Tenenbaum, J. B. (2012). An integrated account of generalization across objects and features. *Cognitive Psychology, 64*, 35-73.

Lee, H. S., & Holyoak, K. J. (2008). The role of causal models in analogical inference. *Journal of Experimental Psychology: Learning, Memory and Cognition, 34*, 1111-1122.

Lombrozo, T. (2007). Simplicity and probability in causal explanation. *Cognition*, *55*, 232-257.

Lu, H., Yuille, A. L., Liljeholm, M., Cheng, P. W., & Holyoak, K. J. (2008). Bayesian generic priors for causal learning. *Psychological Review, 115*, 955-984.

Novick, L. R., & Cheng, P. W. (2004). Assessing interactive causal influence. *Psychological Review, 111*, 455-485.

Oppenheimer, D. M. (2004). Spontaneous discounting of availability in frequency judgment tasks. *Psychological Science, 15*, 100-105.

Rehder, B. (2003a). Categorization as causal reasoning. *Cognitive Science, 27*, 709-748.

Rehder, B. (2003b). A causal-model theory of conceptual representation and categorization. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 29*, 1141-1159.

Rehder, B. (2006). When causality and similarity compete in category-based property induction. *Memory & Cognition, 34*, 3-16.

Rehder, B. (2009). Causal-based property generalization. *Cognitive Science, 33*, 301-343.

Rehder, B., & Burnett, R. C. (2005). Feature inference and the causal structure of object categories. *Cognitive Psychology, 50*, 264-314.

Rehder, B., & Hastie, R. (2001). Causal knowledge and categories: The effects of causal beliefs on categorization, induction, and similarity. *Journal of Experimental Psychology:*

General, 130, 323-360.
Rehder, B. & Kim, S. (2006). How causal knowledge affects classification: A generative theory of categorization. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 32,* 659-683.
Rehder, B. & Kim, S. (2009). Classification as diagnostic reasoning. *Memory & Cognition, 37,* 715-729.
Rehder, B. & Kim, S. (2010). Causal status and coherence in causal-based categorization. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 36,* 1171-1206.
Rogers, T. T., & McClelland, J. L. (2004). Semantic cognition: A parallel distributed processing approach: MIT Press.
Shafto, P., Kemp, C., Bonawitz, E. B., Coley, J. D., & Tenebaum, J. B. (2008). Inductive reasoning about causally transmitted properties. *Cognition, 109*, 175-192.
Sloman, S. A., Love, B. C., & Ahn, W. (1998). Feature centrality and conceptual coherence. *Cognitive Science*, 22, 189-228.
Sloman, S. A., & Lagnado, D. A. (2005). Do we "do"? *Cognitive Science, 29*, 5-39.
Sobel, D. M., Tenenbaum, J. B., & Gopnik, A. (2004). Children's causal inferences from indirect evidence: Backwards blocking and Bayesian reasoning in preschoolers. *Cognitive Science, 28*, 303 -333.
Waldmann, M. R., Holyoak, K. J., & Fratianne, A. (1995). Causal models and the acquisition of category structure. *Journal of Experimental Psychology: General, 124*, 181-206.
Waldmann, M. R., & Hagmayer, Y. (2005). Seeing versus doing: Two modes of accessing causal knowledge. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 31*, 216-227.